Personalitatem Lexicon: Um Léxico em Português Brasileiro para Mineração de traços de Personalidade em Textos

Antonio A. A. Machado[1], Magalí T. Longhi[2], Maria A. S. N. Nunes[1], Thiago A. S. Pardo[3]

[1] Programa de Pós Graduação em Ciência da computação – Universidade Federal de Sergipe (UFS) – Aracaju – SE – Brasil

[2] Programa de Pós-Graduação em Informática na Educação – Universidade Federal do Rio Grande do Sul (UFRGS) – Porto Alegre – RS – Brasil

[3] Núcleo Interinstitucional de Linguística Computacional – Universidade de São Paulo (USP) – São Carlos – SP – Brasil

{aliberteinf,gutanunes}@gmail.com, magali@cpd.ufrgs.br, taspardo@icmc.usp.br

Abstract. *This paper presents initial studies about the correlation of lexical information in Portuguese texts with the Big Five model's psychological features and the NEO-IP facets. In particular, we explore the use of the sentiment classes in the LIWC dictionary. The final purpose is to build or to adapt a lexicon to incorporate not only the sentiment and/or polarity of words, but also their possible personality features.*

Resumo. *Este artigo apresenta estudos iniciais sobre a correlação de informações léxicas em textos em Português com características psicológicas do modelo Big Five e as facetas do NEO-IPIP. Discorre-se, especialmente, sobre o uso das classes de sentimentos do dicionário LIWC. O objetivo final é construir ou adaptar um léxico que incorpore não apenas o sentimento e/ou a polaridade das palavras, mas também suas possíveis características de personalidade.*

1. Introdução

O objetivo deste trabalho é apresentar estudos iniciais desenvolvidos para propor a construção/adaptação de um léxico em português do Brasil que contenha características de personalidade associadas às palavras. Tais características serão baseadas no modelo *Big Five* e nas facetas do NEO-IPIP, como usado em [Nunes 2008].

Esta pesquisa congrega as áreas da Computação Afetiva (CA) e Processamento de Linguagem Natural (PLN) com potencial aplicação na Educação. A CA agrega fatores psicológicos em dispositivos computacionais para reconhecer, modelar, responder às emoções humanas e expressar afetividade por meio da interface computacional [Picard 1997]. As técnicas utilizadas são conhecidas como Mineração de Opinião (*Opinion Mining*), Análise de Sentimento (*Sentiment Analysis*), Análise da Subjetividade (*Subjectivity Analysis*) ou Análise de Julgamento (*Appraisal Extraction*) [Pang & Lee 2008].

DOI: 10.5753/cbie.sbie.2015.1122

Em PLN, há grandes esforços na construção de dicionários e léxicos computacionais. Segundo Specia & Nunes (2004), léxicos são recursos criados geralmente de forma manual para posterior tratamento computacional. São também chamados de bases de dados lexicais (*Lexical Databases*). De especial interesse para este trabalho, há o *Linguistic Inquiry e Word Count* (LIWC), que além do dicionário apresenta uma ferramenta para análise textual ao calcular o grau de utilização das palavras [Pennebaker et al. 2001]. Recentemente, o LIWC (dicionário lexical) foi disponibilizado para o idioma português [Balage Filho et al. 2013].

Em termos educacionais, a *Educational Data Mining* (EDM) é a área de aplicação de mineração de dados que gera informações a partir de contextos educativos. Para [Cummings & Maxwell, 2011], a área ainda é pouco explorada no setor educacional. A análise de opiniões dos alunos sobre a condução de um conteúdo, no *twitter*, por exemplo, pode indicar alterações nas práticas pedagógicas.

Em relação à personalidade, Goldberg (1992) a formaliza por meio de cinco grandes traços definidos no modelo *Big Five*: Abertura, Neuroticismo, Extroversão, Socialização e Realização. O NEO-IPIP [Nunes 2008] é um modelo secundário para identificar as facetas de cada grande traço de personalidade. São elas: para **Abertura,** Fantasia, Estética, Sentimentos, Ações Variadas, Ideias e Valores; para **Neuroticismo**, Ansiedade, Raiva/hostilidade, Depressão, Embaraço/constrangimento, Impulsividade e Vulnerabilidade; para **Extroversão**, Acolhimento, Gregarismo, Assertividade, Atividade, Busca de Sensações e Emoções Positivas; para **Socialização,** Competência, Ordem, Senso de Dever, Esforço por Realizações, Autodisciplina e Ponderação; para **Realização**, Confiança, Franqueza, Altruísmo, Complacência, Modéstia e Sensibilidade.

O restante do trabalho está organizado da seguinte forma: a próxima seção apresenta os trabalhos relacionados, indicando modos de extração de personalidade; a Seção 3 contempla a metodologia e os resultados preliminares; finalmente, na última seção, são feitas algumas considerações em relação ao andamento do trabalho.

2. Trabalhos Relacionados

Sinclair (1966) foi pioneiro nos estudos do léxico, traçando o caminho das pesquisas em Linguística de Corpus. Na década de 80, Ortony et al. (1987) apontaram a importância de um léxico afetivo que contenha palavras que se refiram às emoções, como raiva, tristeza, alegria, orgulho e vergonha, entre outras. Por sua vez, os autores Liu et al. (2003) e Ma et al. (2005) apresentaram o uso de um léxico em trabalhos de reconhecimento de emoções expressas textualmente.

Mais recentemente, Maks & Vossen (2012) apresentou um modelo de léxico para a descrição dos verbos, substantivos e adjetivos que podem ser usados em aplicações como análise de sentimento ou mineração de opinião em textos, enquanto Bandhakavi (2014) propôs um conjunto de métodos para extrair as palavras com conotação afetiva.

Em termos de extração de traços de personalidade, Celli (2012) mostrou estudos de como diferentes personalidades podem ser reconhecidas no *Twitter*, a partir das características do grande fator neuroticismo (do modelo *Big Five).* Já Alam (2014) investigou os traços de personalidade do *Big Five* a partir de dados falados, nos corpus

Speaker Personality Corpus (SPC) e *Personable and Intelligent virtual Agents* (PerSIA).

3. Metodologia e Resultados Preliminares

A questão problema do artigo é como identificar os traços de personalidade de um sujeito a partir da mineração da subjetividade em textos. Para responder à questão de pesquisa, está sendo implementado o *Personalitatem Lexicon*, que contém lexemas de conotação afetiva baseada nos traços de personalidade.

A metodologia utilizada é de uma pesquisa aplicada que contém objetivos exploratórios e descritivos com abordagem qualitativa com caráter bibliográfico e experimental. Ela está dividida em 3 (três) etapas: (1) avaliação do léxico LIWC [Pennebaker et al. 2001], que reúne informações sobre os sentimentos associados às palavras, e de sua versão adaptada para a Língua Portuguesa; (2) construção/adaptação do *Personalitatem Lexicon*; (3) experimentação em uma disciplina de graduação com aplicação dos testes de personalidade[1].

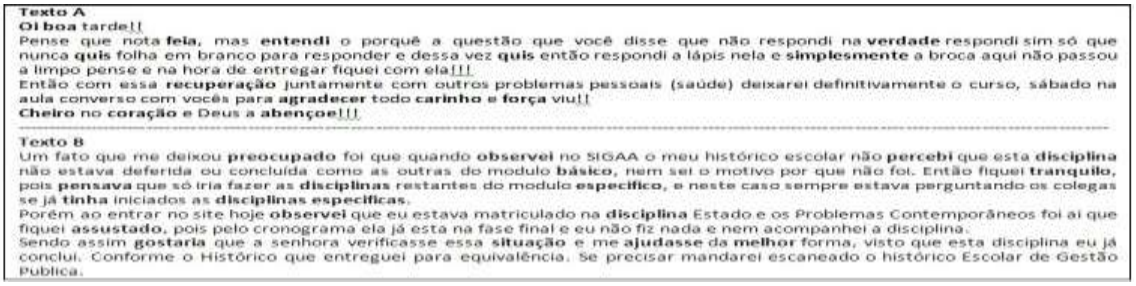
Nesta seção, apresentam-se os resultados parciais do experimento desenvolvido (etapa 3) para examinar os resultados de uma possível adaptação do LIWC do português para conter traços de personalidade ou da construção de um léxico específico.

O experimento se baseou na avaliação de um conjunto de palavras utilizadas em chats de estudantes de um curso de Pós-graduação em Gestão Pública da Universidade Aberta do Brasil (UAB). O julgamento das palavras foi realizado somente por uma pessoa a partir da leitura de duas mensagens disponibilizadas. A análise foi feita a partir da observação das palavras dos textos que estavam presentes no LIWC, em particular, as palavras com polaridades positivas ou negativas, conforme especificado nas classes do LIWC. Isso foi de suma importância, pois, como a pesquisa é inicial, precisava-se de uma referência. Um fator importante foi a dificuldade de se identificar, manualmente, o sentindo destas palavras, pois notou-se ambiguidade durante a análise, o que dificultou na determinação se havia ou não conotação afetiva.

Assim, as palavras afetivas, conforme se apresenta na Figura 1, receberam uma marcação (palavras em negrito). Para toda palavra considerada afetiva (*a*), foi verificado a sua significância em cada grande traço de personalidade. Para isso, foi aplicada uma escala de significação ou peso (P), atribuído pelo pesquisador, conforme a sua relação com o traço de personalidade (não significativa, pouco significativa, significativa, muito significativa). A partir disso, foi aplicado um cálculo para determinar o traço de personalidade (TP) mais expressivo no texto. Então, o traço de personalidade (1-Abertura, 2-Neuroticismo, 3-Extroversão, 4-Socialização, e 5-Realização) é verificado através do somatório das palavras afetivas (associadas com o seu peso) para o traço de análise, dividido pela quantidade de palavras afetivas (W) encontradas no texto. O cálculo é demonstrado pela fórmula:

$$\mathrm{TP}_{(1..5)} = \frac{\sum_{i=1}^{n} a_i * P}{w}$$

[1] Portal do Personalitatem: http://personalitatem.ufs.br/



Texto A

Oi boa tarde!!!
Pense que nota **feia**, mas **entendi** o porquê a questão que você disse que não respondi na **verdade** respondi sim só que nunca **quis** folha em branco para responder e dessa vez **quis** então respondi a lápis nela e **simplesmente** a broca aqui não passou a limpo pense e na hora de entregar fiquei com ela!!!
Então com essa **recuperação** juntamente com outros problemas pessoais (saúde) deixarei definitivamente o curso, sábado na aula converso com vocês para **agradecer** todo **carinho** e **força** viu!!!
Cheiro no **coração** e Deus a **abençoe**!!!

Texto B

Um fato que me deixou **preocupado** foi que quando **observei** no SIGAA o meu histórico escolar não **percebi** que esta **disciplina** não estava deferida ou concluída como as outras do modulo **básico**, nem sei o motivo por que não foi. Então fiquei **tranquilo**, pois **pensava** que só iria fazer as **disciplinas** restantes do modulo **específico**, e neste caso sempre estava perguntando os colegas se já **tinha** iniciados as **disciplinas específicas**.
Porém ao entrar no site hoje **observei** que eu estava matriculado na **disciplina** Estado e os Problemas Contemporâneos foi aí que fiquei **assustado**, pois pelo cronograma ela já esta na fase final e eu não fiz nada e nem acompanhei a disciplina.
Sendo assim **gostaria** que a senhora verificasse essa **situação** e me **ajudasse** da **melhor** forma, visto que esta disciplina eu já conclui. Conforme o Histórico que entreguei para equivalência. Se precisar mandarei escaneado o histórico Escolar de Gestão Publica.

Figura 1. Mensagens que expressam estados afetivos (os grifos são nossos)

Analisando os textos, podemos concluir que o traço de personalidade mais provável para o sujeito do Texto A é o de Socialização (do modelo *Big Five*). Já no Texto B, o sujeito tende para realização. A relação das palavras com os traços de personalidade e pesos atribuídos manualmente é apresentada na Tabela 1 enquanto os resultados dos cálculos efetuados pela fórmula do TP são demonstrados na Tabela 2.

Tabela 1. Relação das palavras com os traços de personalidade e pesos

Sujeito 1	Palavras Texto A															
Big Five	Oi	Boa	Feia	Entendi	Verdade	Quis	Quis	Simplesm*	Recuper*	Agradecer	Carinho	Força	Cheiro	Coração	Abençoe
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	3	0	0	1	1	0	0	0	0	0	0	0	0
E	3	3	0	0	0	0	0	0	0	3	3	0	3	3	3
S	3	0	0	1	3	1	1	2	0	2	2	0	2	3	3
R	2	2	2	2	2	0	0	2	2	2	0	2	0	0	0

Sujeito 2	Palavras Texto B													
Big Five	Preocup*	Observ* (2x)	Percebi	Disciplina (4x)	Básico	Tranqui*	Pensava	Específico (2x)	Tinha	Assustado	Gostaria	Situação	Ajudasse	Melhor
A	0	1	1	0	0	0	1	0	1	0	0	2	0	0
N	2	0	0	0	0	0	0	0	0	3	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	1	
R	1	0	1	2	1	3	0	2	0	0	2	0	2	2

Legenda: A=Abertura; N=Neuroticismo; E=Extroversão; S=Socialização; R=Realização; 0=Nada Significativo; 1=Pouco Significativo; 2= Significativo; 3=Muito Significativo.

Tabela 2. Resultados da inferência dos traços de personalidade (TPs) em texto

Sujeito	Textos	A	N	E	S	R
1	A	0%	8%	31%	33%	28%
2	B	13%	13%	0%	15%	59%

Os valores apresentados na Tabela 2 são resultantes dos cálculos efetuados pela fórmula anterior (TPs). Nessa análise preliminar dos *chats* de uma disciplina do curso [xxx(omitido)xxx] da UAB, vê-se que há palavras que indicam os traços de personalidade, mas que ainda não é possível distingui-los apropriadamente. Em trabalhos futuros, com mais dados, essa modelagem deve ser aperfeiçoada e estendida, permitindo uma aferição melhor dos traços de personalidade.

4. Considerações Finais

Apresentou-se, neste artigo, uma tentativa inicial de se mapear a ocorrência de palavras indicadoras de personalidade em textos de estudantes. Ressalta-se que os traços de personalidade são características duradouras no ser humano. A avaliação de um (ou poucos) textos, necessariamente, pode não refletir os fatores psicológicos. Assim, é preciso buscar vários textos de um sujeito em um período maior possível para poder inferir com mais acuidade os traços de personalidade.

Referências Bibliográficas

Alam, F & Riccardi, G. (2014) Fusion of Acoustic, Linguistic and Psycholinguistic Features for Speaker Personality Traits Recognition. In: Acoustics, Speech and Signal Processing (Icassp), Ieee International Conference, pp. 955-959.

Balage Filho, P.P.; Aluísio, S.M.; Pardo, T.A.S. (2013). An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. In the Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology – STIL, pp. 215-219. October 21-23, Fortaleza/Brazil.

Bandhakavi, A. & Massie, S. (2014) Generating A Word-Emotion Lexicon From # Emotional Tweets, pp. 12–21.

Celli, F & Rossi, L. (2012) The Role of Emotional Stability In Twitter Conversations. In: Proceedings of The Workshop On Semantic Analysis In Social Media. Association For Computational Linguistics, pp. 10-17.

Cummings, Richard G., and Maxwell Hsu. (2011) The effects of student response systems on performance and satisfaction: An investigation in a tax accounting class. Journal of College Teaching and Learning (TLC) 4.12.

Goldberg, L., (1992) R. The Development of Markers for The Big Five Factor Structure. In Psychological Assessment, 4(1), pp. 26–42.

Liu, H.; Lieberman H. & Selker T. (2003) A Model of Textual Affect Sensing Using Real-World Knowledge. Proceedings of The 8th International Conference on Intelligent User Interfaces, 12-15, Miami, Florida, Usa.

Ma, C.; Prendinger, H. & Ishizuka, M. (2005) Emotion Estimation and Reasoning Based on Affective Textual Interaction, In Affective Computing and Intelligent Interaction. (First Int'l Conf. Acii. pp.622-628. Beijing, China.

Maks, I. & Vossen, P. (2012) A Lexicon Model For Deep Sentiment Analysis and Opinion Mining Applications. Decision Support Systems, Elsevier, V. 53, N. 4, pp. 680–688.

Nunes, M. A. S. N. (2008) Recommender System based on Personality Traits. Universite Montpellier 2-Lirmm-França.

Ortony, A.; Clore, G. & Colins, A. (1998) The Cognitive Structure of Emotions, Cambridge University Press.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends In Information Retrieval 2(1-2):1-135.

Pennebaker, J. W.; Francis, M. E. & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates.

Picard, R. W. (1997) Affective Computing. Mit Press, Cambridge, Ma, Usa.

Sinclair, J. (1996) Eagles Preliminary Recommendations On Corpus Typology Eag--Tcwg--Ctyp/P Version of May, ILC-CNR, Pisa.

Specia, L. & Nunes, M. (2004) Desambiguação Lexical Automática de Sentido: Um Panorama. Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional. NILC - ICMC-USP, São Carlos, SP, Brasil.

Schultz, D. (1990) Theories of Personality. 4ª Ed. Brooks/Cole.